Advancements in the human genome reference assembly (GRCh38)

Tayebeh Rezaie, Ph.D.
NCBI

16 October 2019

Genome Reference Consortium

NCBI

McDONNELL GENOME INSTITUTE

wellcome sanger institute

EMBL-EBI

GRC

- Valerie Schneider
- Kerstin Howe
- Tina Graves
- Paul Flicek
- Tayebeh Rezaie
- Nathan Bouk
- Hsiu-Chuan Chen
- Jo Wood
- Joanna Collins
- Sarah Pelan
- Will Chow
- James Torrance
- Ying Sims
- Derek Albracht
- Milinn Kremitzki

Thanks to many GRC Collaborators
https://www.ncbi.nlm.nih.gov/grc/credits/

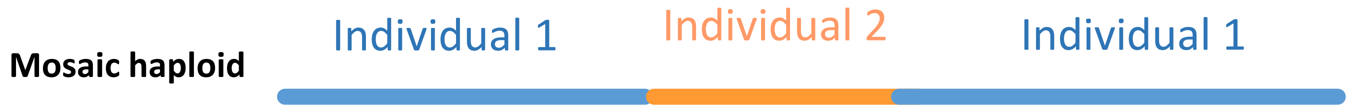
Funding:

- This work was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.
- The European Molecular Biology Laboratory.
- The Wellcome Trust, UK.
- The MGI was supported by National Institutes of Health grants 5U54HG003079, 5U41HG007635 and 5U24HG009081.

History of reference assembly

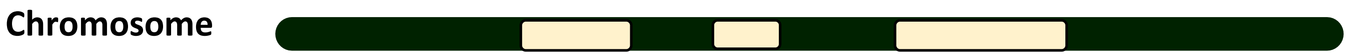
GRCh38/Reference genome:

- A critical resource to the basic & clinical research community, coordinate system, annotation source & discovery of disease-associated variants
- Sanger seq. clone-based from **H**uman **G**enome **P**roject; multiple individuals

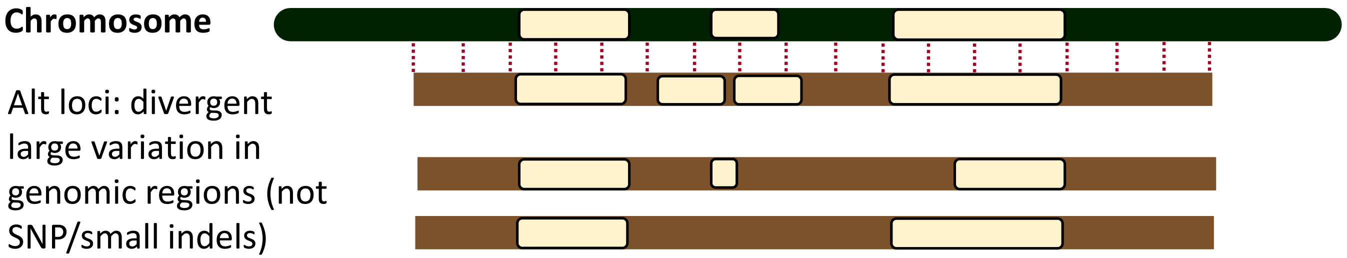


HGP → GRC: reference maintaining, improving and updates

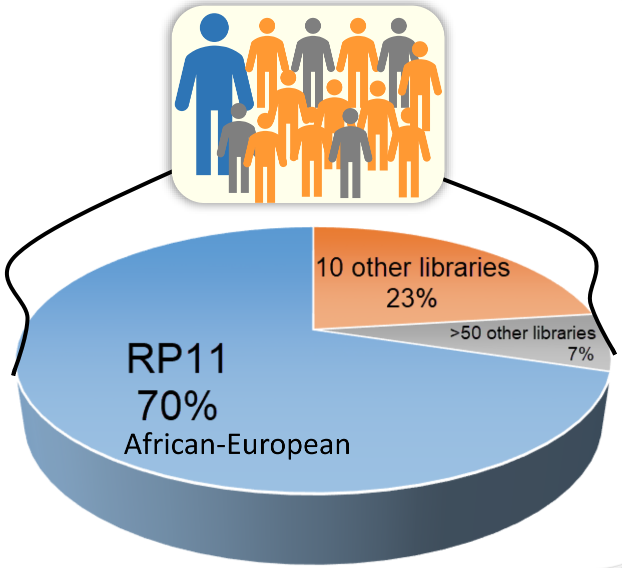
HGP model (2003): each genomic region was represented with one sequence

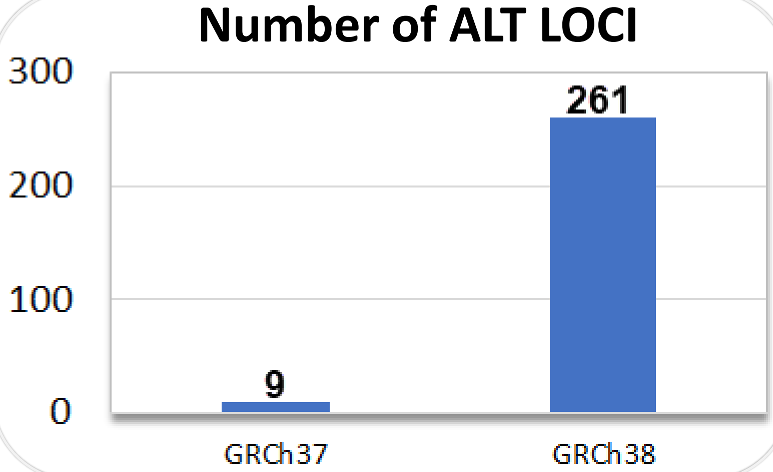


Current model: ALT LOCI added to represent population genomic diversity



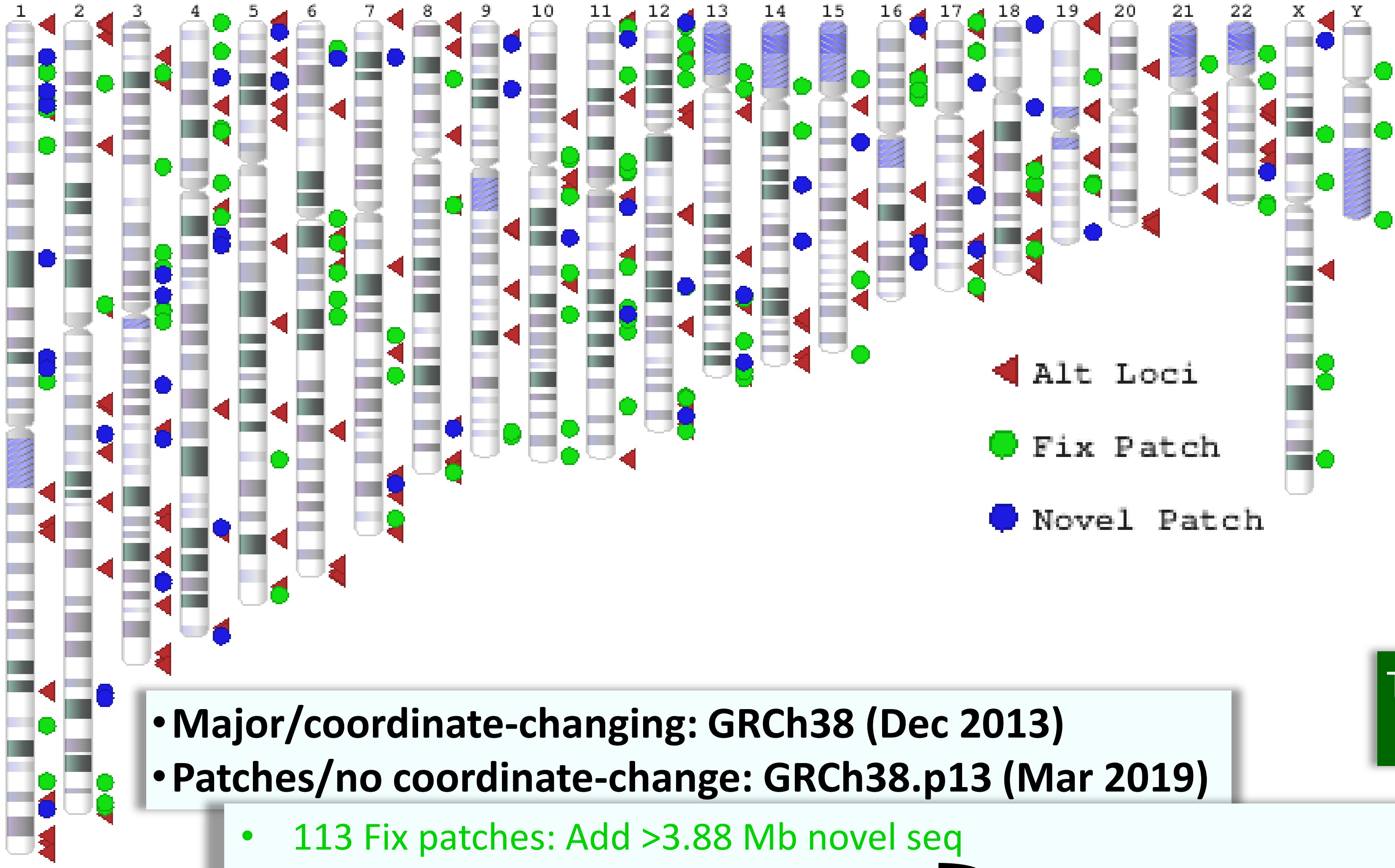
Alt loci: divergent large variation in genomic regions (not SNP/small indels)

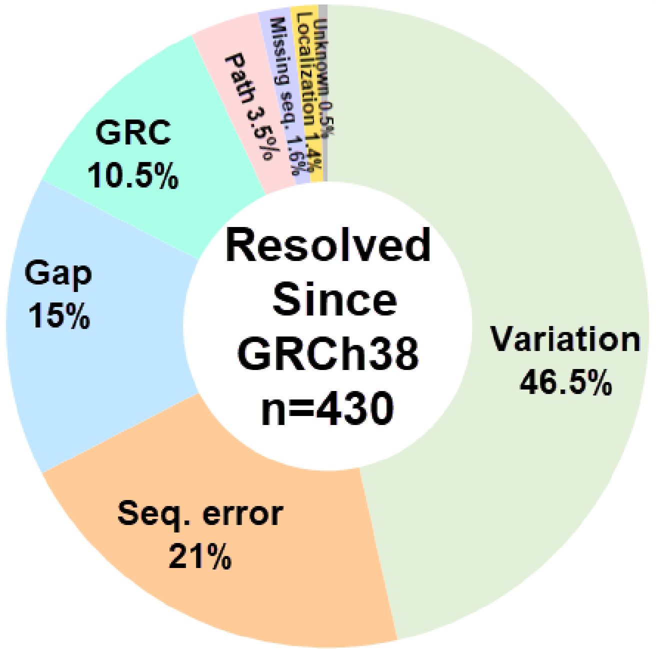




Genome Reference Consortium

Reference assembly updates





Reference updates released as patches: 185/430 (42%)

The new version of the reference should capture ALL the updates to GRCh38

- **Major/coordinate-changing: GRCh38 (Dec 2013)**
- **Patches/no coordinate-change: GRCh38.p13 (Mar 2019)**

- 113 Fix patches: Add >3.88 Mb novel seq
- 72 Novel patches: Add >1.1 Mb novel seq
- 261 ALT Loci: Add 3.6 Mb novel seq

The notion for variant representation has started long time ago.

Genome Reference Consortium

Data provided by Tina Graves

Curation of reference assembly

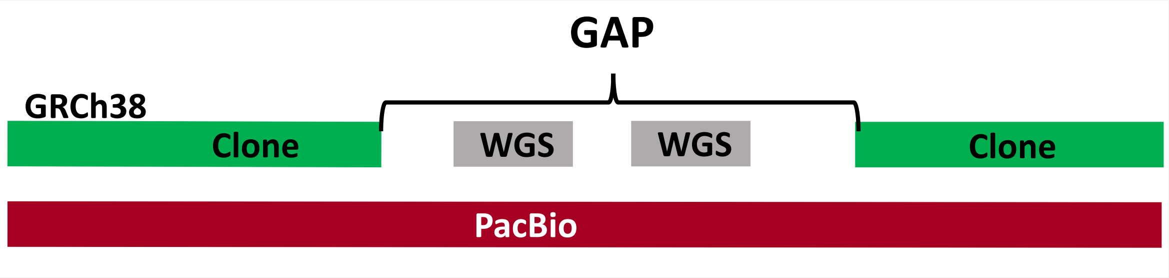
- Issue sources: GRC assembly evaluation, reports from collaborators, community, literature
- Technology: sequencing, FISH, Optical Mapping
- Data resources: sequences generated by GRC or available in public database (clones, WGS, PCR products)

Evaluation of gaps in GRCh38

- Gap count = 196

Excluded biological gaps & gaps within WGS scaffolds

- Reports of new assm that can close ref. gaps
- To identify gaps that can be spanned



Alignments of 8 diploid PacBio assemblies to the reference:

- Spanned with the same amount of seq: 26 (missing seq.)
- Spanned with varying amount of seq: 3 (variation)
- Spanned by some not all assemblies: 24 (complex, missing + variation)
- The remaining gaps are under review

https://www.genome.wustl.edu/research/projects/

Curation of reference assembly: Missing sequences

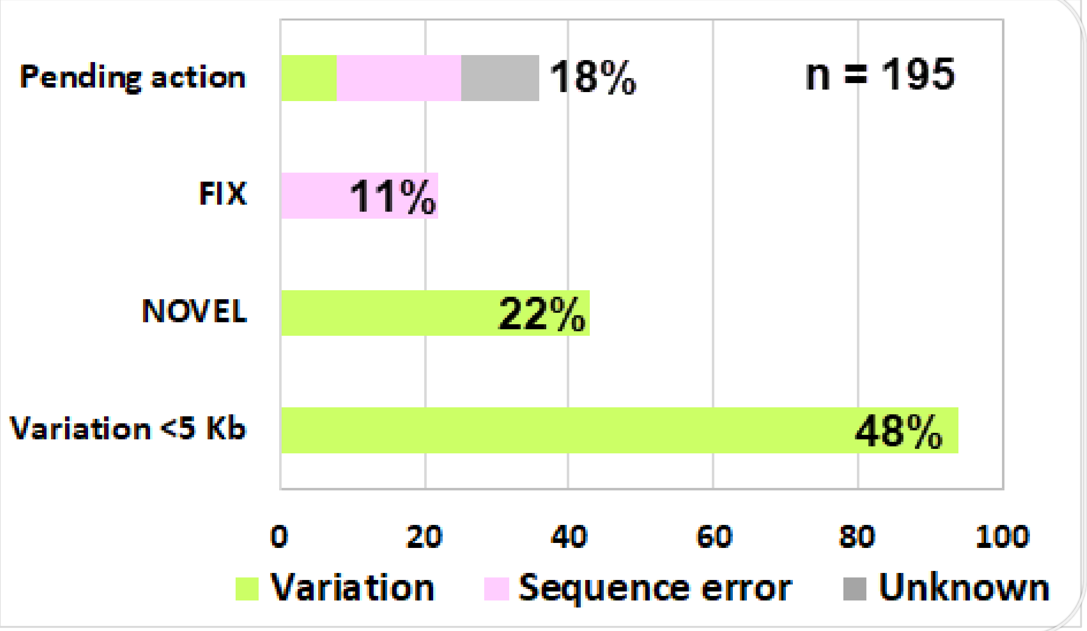
Evaluation to distinguish error vs. variation
→ Find chr. context for missing seq.
→ Add variants (>5 kb) as novel patches

Data sources:

- Eichler’s lab (Kidd et al. (2010) PMID: 20440878), structurally variant fosmid seq.
- Heng Li (GCA_000786075.2), a set of non-redundant seq. absent in GRCh38 and ALTs

Reported genome issues = 195

- Resolved no change: 94 (variation < 5 Kb)
- Patches (started adding from p1 in 2014)
  - FIX = 22
  - NOVEL = 43
- Pending action: 36 (Variation 8, sequence error 17, Unknown 11)



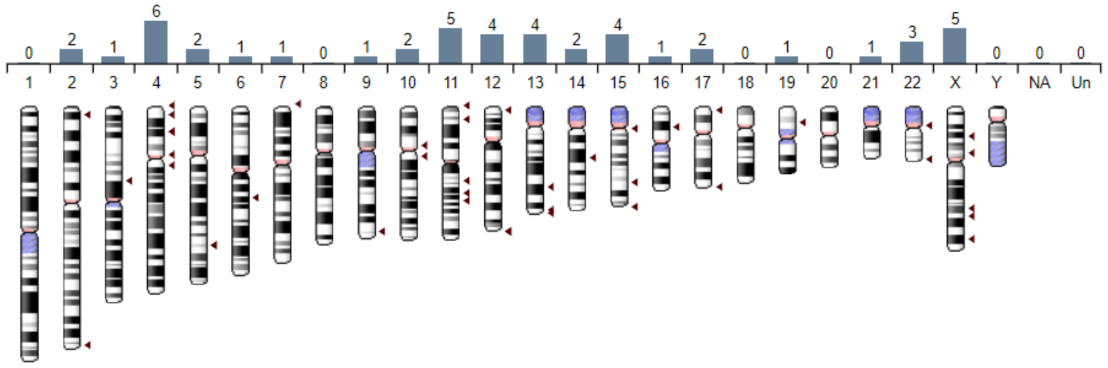
GRCh38.p13 updates to reference assembly

The most recent curation to GRCh38:

- FIX patches (43) + NOVEL (2)
- Added >0.5 Mb novel sequence
  - Gap closure: 28
  - Seq. error correction: 8
  - Path: 2
  - For p-arm of acrocentric chrs: 5

- **Highlights of p13:**
  - Improved clinically important genomic regions
    - Prader-Willi (5.5 Mb, 1.63 Mb unique)
    - CT47A gene cluster
  - Improved gene representations: SLC5A11, GCNT2, SAMD1, GRCK1, C1R, ECSCR, 5S rRNA

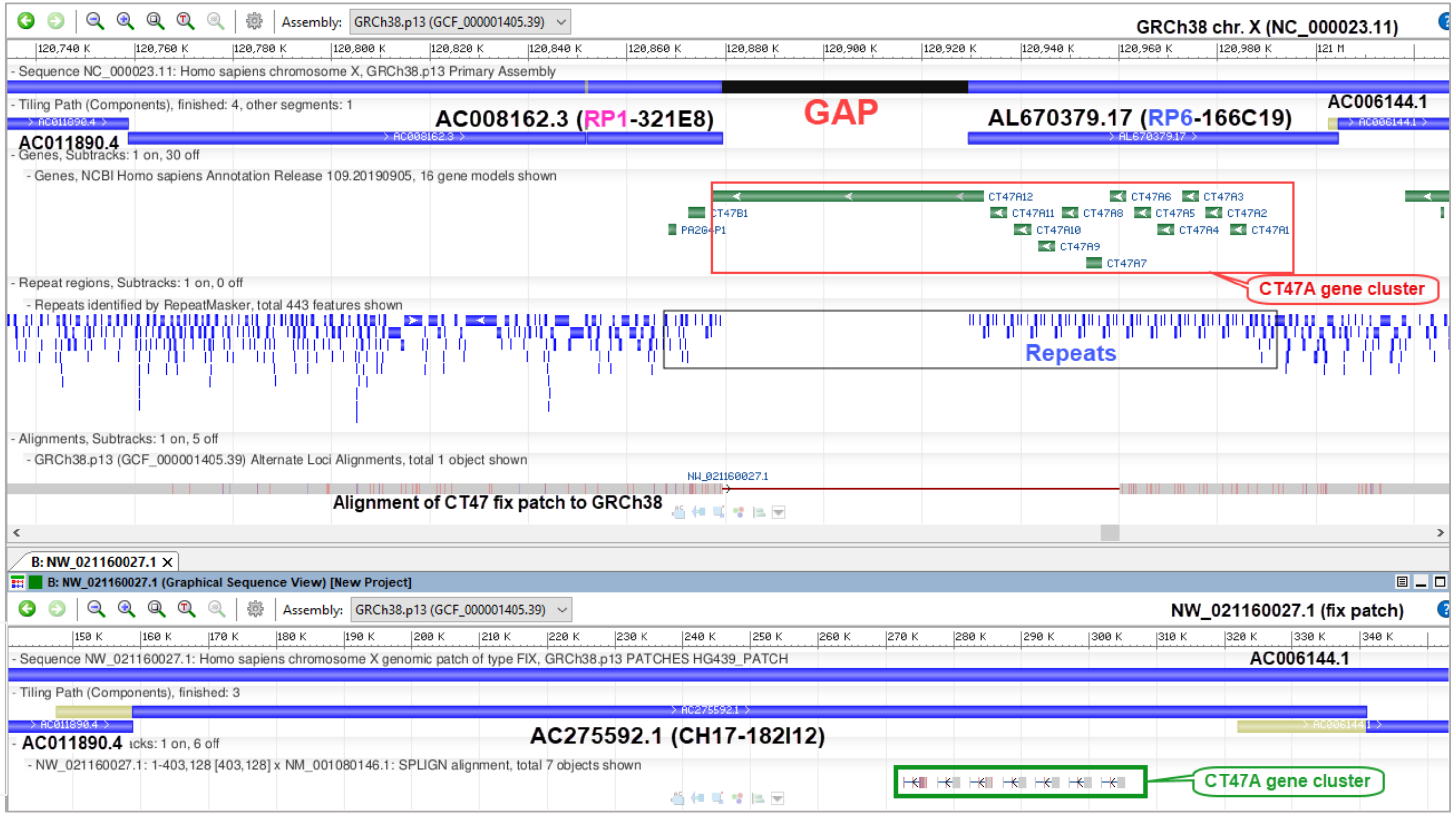
Chr. distribution of GRCh38.p13 patches



- Sequence data sources for updates:
  - CHM1 assm: 21
  - CHM13 assm: 12
  - Other WGS assm: 3
  - Clones: 9

Correction of an assembly false gap caused by haplotype incompatibility

Mix haplotype representation of CT47A in GRCh38
Long haplotype: 12 copies



Assembly: GRCh38.p13 (GCF_000001405.39)

GRCh38 chr. X (NC_000023.11)

- Sequence NC_000023.11: Homo sapiens chromosome X, GRCh38.p13 Primary Assembly
- Tiling Path (Components), finished: 4, other segments: 1
- Genes, Subtracks: 1 on, 30 off
  - Genes, NCBI Homo sapiens Annotation Release 109.20190905, 16 gene models shown
- Repeat regions, Subtracks: 1 on, 0 off
  - Repeats identified by RepeatMasker, total 443 features shown
- Alignments, Subtracks: 1 on, 5 off
  - GRCh38.p13 (GCF_000001405.39) Alternate Loci Alignments, total 1 object shown

Alignment of CT47 fix patch to GRCh38

B: NW_021160027.1 (Graphical Sequence View) [New Project]

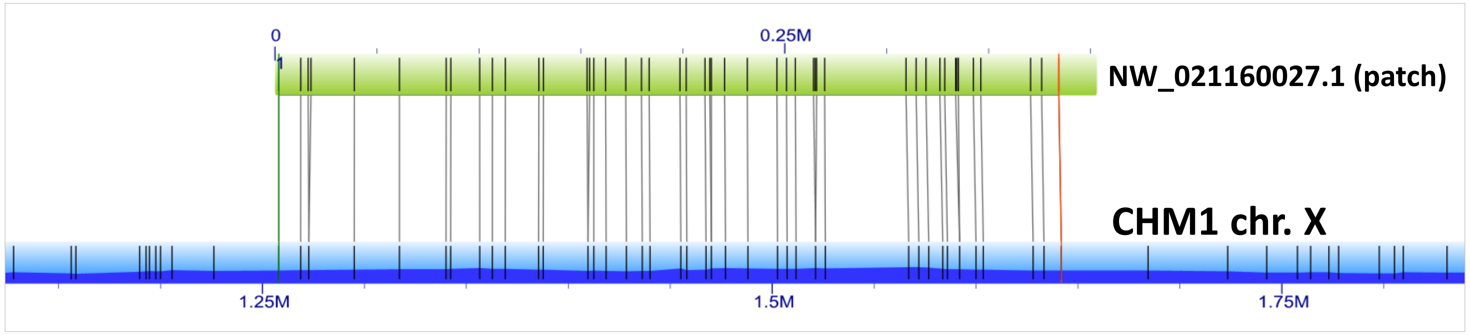
Assembly: GRCh38.p13 (GCF_000001405.39)

NW_021160027.1 (fix patch)

- Sequence NW_021160027.1: Homo sapiens chromosome X genomic patch of type FIX, GRCh38.p13 PATCHES HG439_PATCH
- Tiling Path (Components), finished: 3
- [...]icks: 1 on, 6 off
  - NW_021160027.1: 1-403,128 [403,128] x NM_001080146.1: SPLIGN alignment, total 7 objects shown

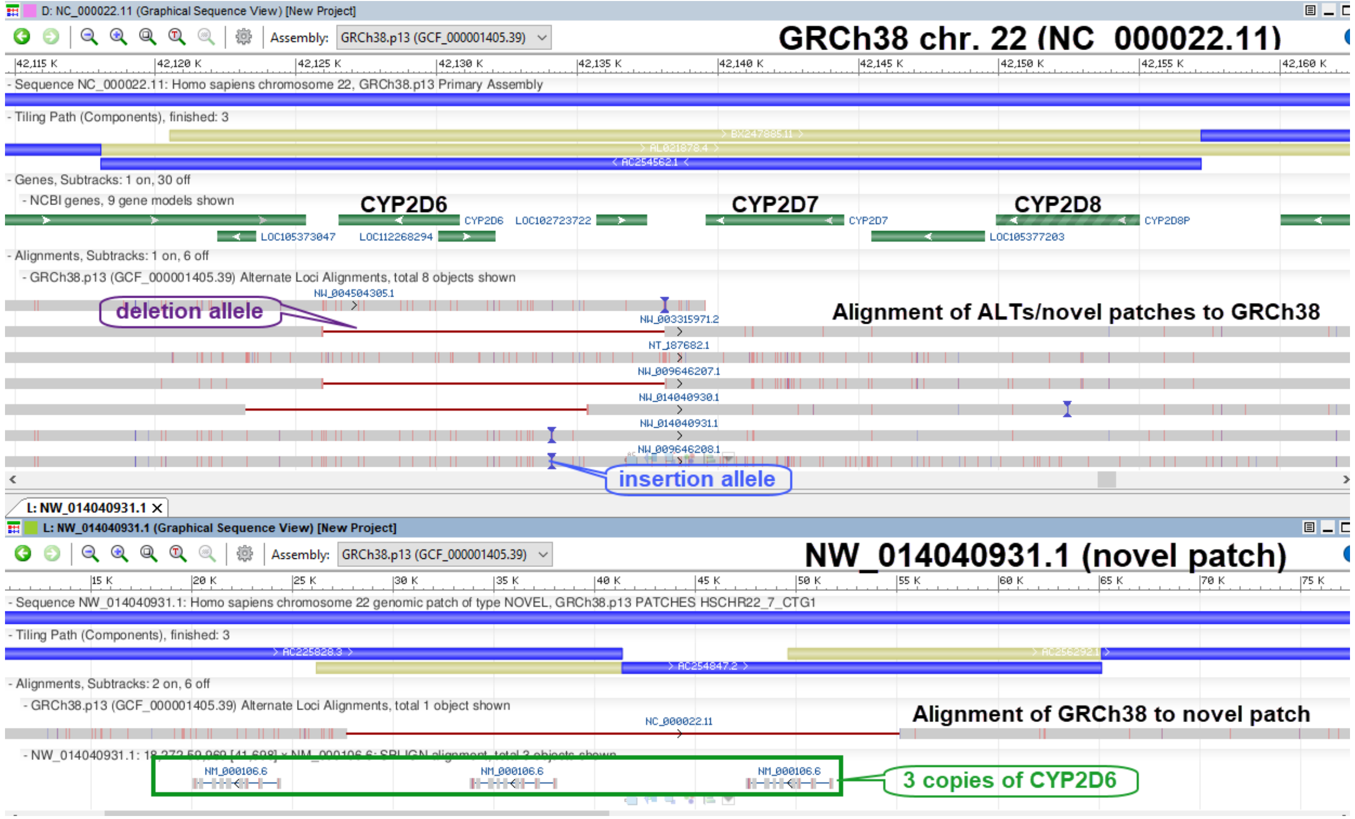
Single haplotype representation of CT47A in GRCh38.p13
Short haplotype: 7 copies

CHM1 Optical Map supporting the updated CT47A haplotype



CYP2D6 haplotypes: genomic diversity of a clinically important region Involved in metabolizing many prescribed drugs

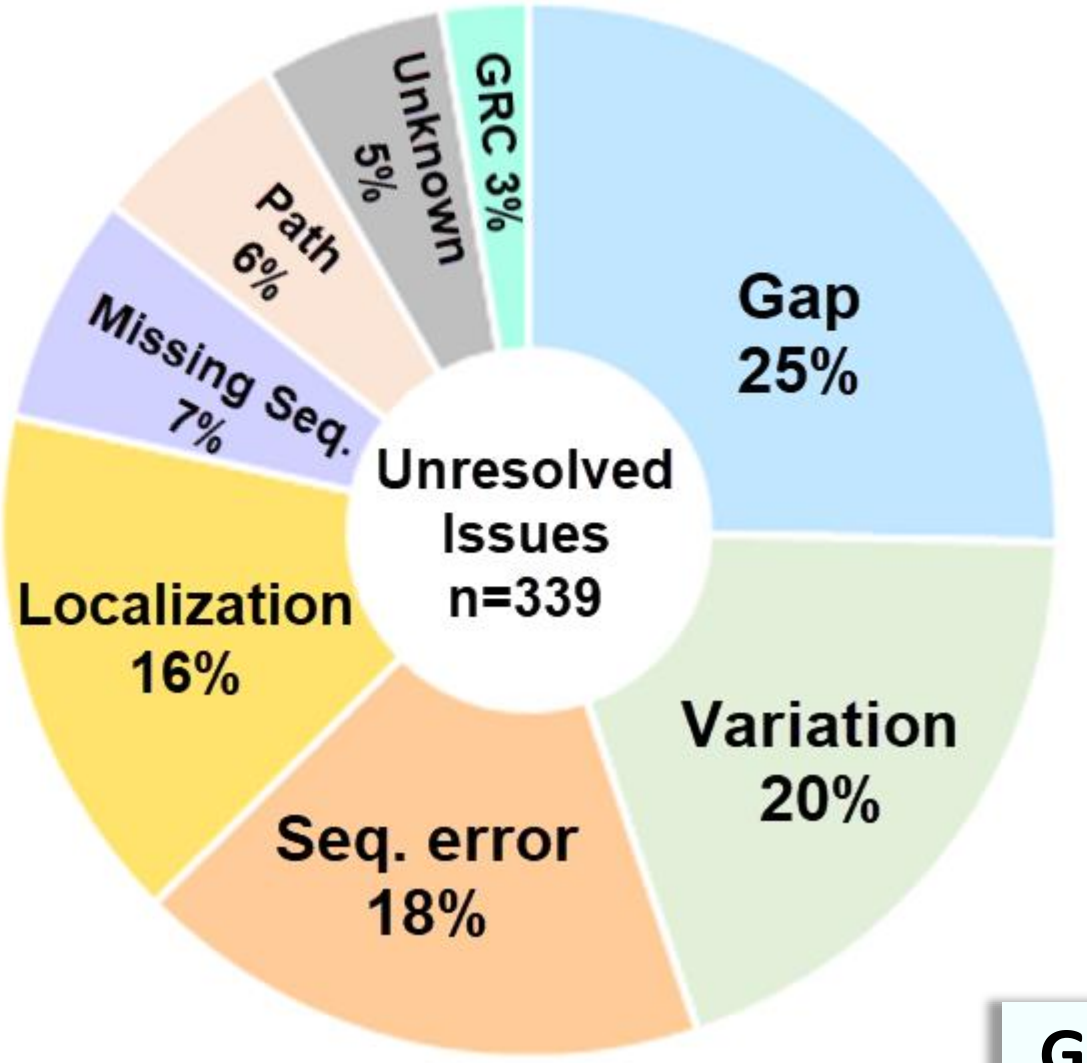
Alignment of alt loci and patch scaffolds to the CYP2D6 region of chr. 22


D: NC_000022.11 (Graphical Sequence View) [New Project]
Assembly: GRCh38.p13 (GCF_000001405.39)
GRCh38 chr. 22 (NC_000022.11)
42,115 K
42,120 K
42,125 K
42,130 K
42,135 K
42,140 K
42,145 K
42,150 K
42,155 K
42,160 K
Sequence NC_000022.11: Homo sapiens chromosome 22, GRCh38.p13 Primary Assembly
Tiling Path (Components), finished: 3
Genes, Subtracks: 1 on, 30 off
NCBI genes, 9 gene models shown
CYP2D6
CYP2D7
CYP2D8
CYP2D6
LOC102723722
CYP2D7
CYP2D8P
LOC105373047
LOC112268294
LOC105377203
Alignments, Subtracks: 1 on, 6 off
GRCh38.p13 (GCF_000001405.39) Alternate Loci Alignments, total 8 objects shown
NW_004504305.1
deletion allele
NW_003315971.2
Alignment of ALTs/novel patches to GRCh38
NT_187682.1
NW_009646207.1
NW_014040930.1
NW_014040931.1
NW_009646208.1
insertion allele
L: NW_014040931.1
L: NW_014040931.1 (Graphical Sequence View) [New Project]
Assembly: GRCh38.p13 (GCF_000001405.39)
NW_014040931.1 (novel patch)
15 K
20 K
25 K
30 K
35 K
40 K
45 K
50 K
55 K
60 K
65 K
70 K
75 K
Sequence NW_014040931.1: Homo sapiens chromosome 22 genomic patch of type NOVEL, GRCh38.p13 PATCHES HSCHR22_7_CTG1
Tiling Path (Components), finished: 3
Alignments, Subtracks: 2 on, 6 off
GRCh38.p13 (GCF_000001405.39) Alternate Loci Alignments, total 1 object shown
NC_000022.11
Alignment of GRCh38 to novel patch
NM_000106.6
NM_000106.6
NM_000106.6
3 copies of CYP2D6

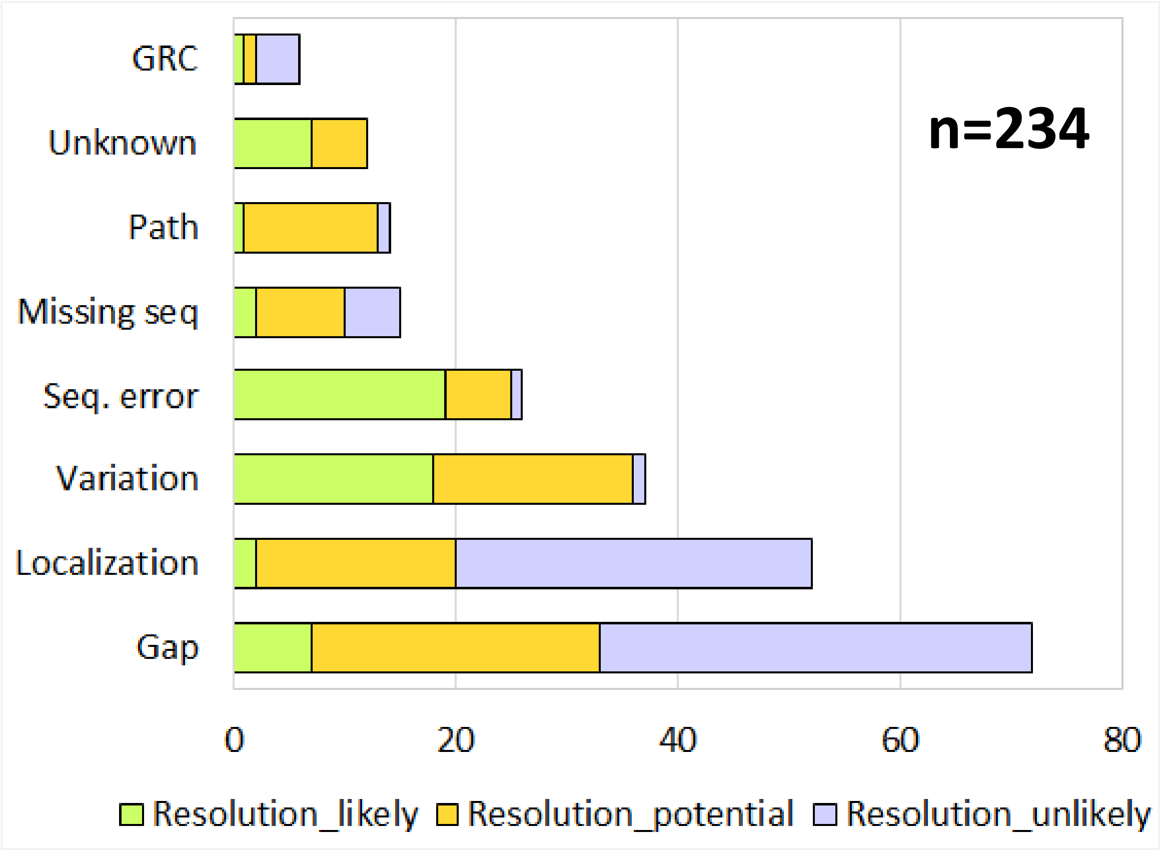
Scaffolds providing alternate sequence representations of CYP2D6 region

GenBank Acc.	RefSeq Acc.	Population	CYP2D6	CYP2D7	CYP2D8
KN196485.1	NW_009646207.1	African	Deletion	Single Copy	Single Copy
KB663609.1	NW_004504305.1	African	Duplication	Single Copy	Single Copy
KN196486.1	NW_009646208.1	East Asian	Duplication	Single Copy	Single Copy
KQ458387.1	NW_014040930.1	East Asian	Deletion	Single Copy	Duplication
KQ458388.1	NW_014040931.1	East Asian	3 Copies	Single Copy	Single Copy
KQ759761.1	NW_015148968.1	European	Single Copy	Duplication	Single Copy
GL383582.2	NW_003315971.2	Unknown	Deletion	Single Copy	Single Copy

Unresolved genome issues



Current curation status



Resolution likelihoods as determined by the GRC review

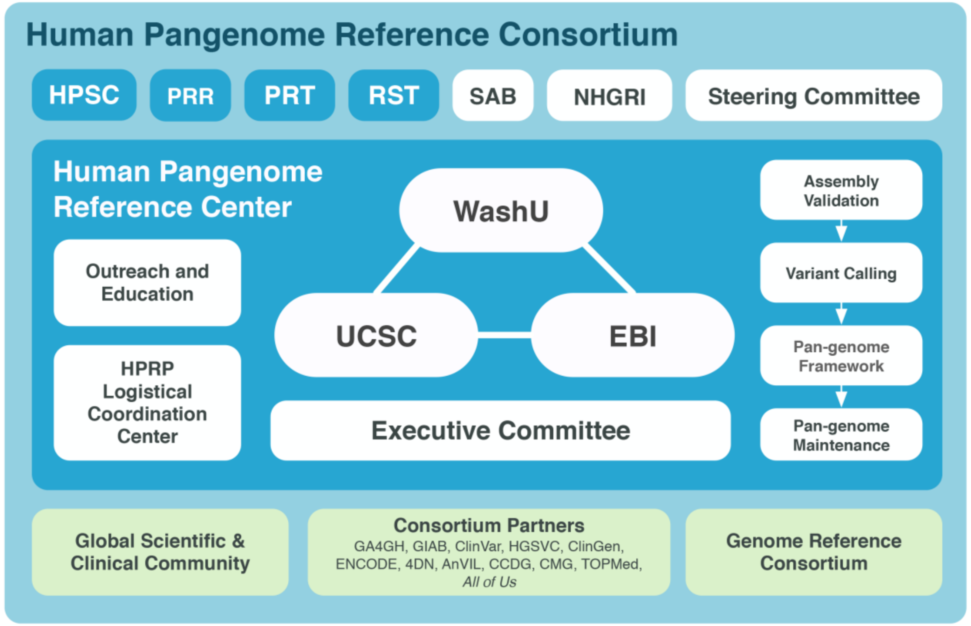
GRCh38.p14 is planned for release in 2020

Conclusion and Future

GRCh38.p14: coming in 2020

The reference has informed its own evolution.

Washington Post


Human Pangenome Reference Consortium

HPSC | PRR | PRT | RST | SAB | NHGRI | Steering Committee

Human Pangenome Reference Center

Outreach and Education

HPRP Logistical Coordination Center

WashU

UCSC

EBI

Executive Committee

Assembly Validation → Variant Calling → Pan-genome Framework → Pan-genome Maintenance

Global Scientific & Clinical Community

Consortium Partners
GA4GH, GIAB, ClinVar, HGSVC, ClinGen, ENCODE, 4DN, AnVIL, CCDG, CMG, TOPMed, *All of Us*

Genome Reference Consortium

MGI, a GRC member, has been awarded by NHGRI to:

- Produce 350 whole genome phased diploid assm.;
- Identify SVs between samples and current GRCh38;
- Incorporate those SVs into the Reference, likely as a graph representation.

GRCh39 is pending. The GRC is engaged in validation, providing curation tools and support to the pan-genome assemblies.